# Systemic similarity analysis of compatibility drug-induced multiple pathway patterns in vivo

Zhong Wang[1*], Zhan-jun Zhang[2], Yi Wang[3], Yiyu Cheng[3], Weida Tong[4], Yongyan Wang[1]

[1]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences**,** 18 Baixincang, Dongzhimennei, Beijing 100700,China; [2]Beijing Normal University, The Key Laboratory of Traditional Chinese Medicine Protection and Utilization, 19 XinJieKouWai Street, HaiDian District, Beijing, PA 100875, China; [3]Pharmaceutical Research Institute, Zhejiang University; [4]National Center for Toxicological Research(NCTR),FDA,3900 NCTR Road,Jefferson,AR 72079,USA

*Correspondence: Fax: +86-10-62874049,E-mail: zhonw@vip.sina.com

A major challenge in post-genomic research is to understand how physiological and pathological phenotypes arise from the networks of expressed genes[1] and to develop powerful tools for translating the information exchanged between gene and the organ system networks. Although different expression modules may contribute independently to different phenotypes[2,3], it is difficult to interpret microarray experimental results at the level of single gene associations[4]. The global effects[5] and response pathways[6] of small molecules in cells have been investigated, but the quantitative details of the activation mechanisms of multiple pathways in vivo are not well understood. Similar response networks[7,8,9] indicate similar modes of action[10], and gene networks may appear to be similar despite differences in the behaviour of individual gene groups[11,12]. Here we establish the method for assessing global effect spectra of the complex signaling forms using Global Similarity Index (GSI) in cosines vector included angle. Our approach provides quantitative multidimensional measures of genes expression profile based on drug-dependent phenotypic alteration in vivo. These results make a starting point for identifying relationships between GSI at the molecular level and a step toward phenotypic outcomes at a system level to predict action of unknown compounds and any combination therapy.

We applied both top-down and bottom-up approaches in this research

on the ischaemic mouse hippocampus (Fig.1). Baicalin (BA), jasminoidin (JA), cholic acid (CA), Concha margaritifera (CM) and nimodipine (NI) [13,14] have different effects on neurons subjected to an ischaemic insult. IV after occlusion of the middle cerebral artery was significantly smaller after treatment with each compound except CM; the reduction in IV was greatest for NI (Fig.2a, F=17.01, $*P < 0.05$, $**P < 0.01$ vs. vehicle; mean ± SD, ANOVA, n = 7–9). In the vehicle- and sham-treated mice, 24.06% (90/374) of genes differed in their expression in the hippocampus. Although 90 genes exhibited highly significant differences in expression, only one-third (32 genes) of the cDNAs showed average differences greater than twofold (Fig. 2b). Compared with the vehicle-treated condition, expression differed in 99, 72, 106, 62 and 62 genes in ischaemic mouse hippocampi treated with BA, JA, CA, CM, and NI, respectively. Only eight genes overlapped in the hippocampi treated with BA, JA, CA, and NI versus CM and vehicle, and 19 genes were shared between all groups (Supp Table 1) (Fig. 2b). Based on data from a previous study[15], these genes were expected to respond to ischaemic injury or are related to hippocampal function, except that Selenbp2 was null. An interesting contradiction is that the eight overlapping genes should contribute most to the pharmacological effect, but this was not supported by direct comparison in PCA (Fig.2c), which showed that PCA1

accounted for 56.3% and PCA1-2 for 67.8% of the pharmacological variation. This dominant pattern of expression of the top five major genes was clearly associated with G protein-coupled receptors (GPCR**s**) and Ras transcription, which did not overlap in the larger altered genes that protect against cerebral ischaemia such as RGS6, Cbx3, and Grb2 (Fig.2c)[16]. Only three overlapping genes in black pane (Fig.2d) contributed to the pharmacological effect based on the changed phenotypes, although 15 genes were shared in all compound groups (Suppl Table 2). The PCA showed that another five top genes might contribute to the pharmacological action, of which three genes Rgs6, Cbx3, and Grb2 in black pane were consistent with the above analysis (Fig.2c). PCA1 accounted for 60.29% and PCA2 14.50% of the variation. We infer that studying the effects of these candidate genes on pharmacological behaviour requires more than direct comparison analysis. Expressions of the eight selected genes observed by microarray analysis were independently confirmed by Real-time RT-PCR, which showed that Rgs6 had the highest activity (Fig.2e).

We next designed two two-level factional experiments (Fig.3a-d) to study the mechanism of the concept of multivalent chemotherapy[17]. The ratios of overlapping genes (ROG) (23/133) and non-overlapping to overlapping genes (RNOG) (23/23/133) were 0.1729 and 4.78, respectively.

The ROG and RNOG of JA+BA versus BA or JA alone were 0.2667 (44/165) and 2.75 (44/44/165), respectively (Fig.3c), suggesting the presence of more overlapping genes in the combination treatment (Fig.3b-d). Variations of gene expression profiles of eight pathways (Fig.3e) revealed that the extracellular signal-regulated kinase mitogen-activated protein kinase (Erk-MAPK) network shared different conditions in these treated groups because only some significantly altered genes emerged in this network (Fig.3f). One challenge that emerged was whether direct comparison analysis of the ROG and NROG could sufficiently reveal essential information about broad changes in so many pathways.

Some differences in gene expression of less than 1.5-fold are robustly associated with behavioural differences[3] and might be as important as those of genes with greater differences in expression condition. The fold change in expression may not be linearly related to phenotype behaviour because a smaller fold change (–1.72) had a higher correlation coefficient (0.95), and a larger fold change (–2.83) had a lower coefficient (0.85)[18]. We hypothesized that GSI could be used to quantitatively analyse the gene expression pattern of mouse hippocampus treated with BA, JA, CA, or NI alone or in combination. The GSI from this approach was greater than the Pearson coefficients, which were all < 0.7, and near to the Euclidean distance (the

range not wider than our approach) (Fig.4a).

GSI after treatment with BA, JA, CA, or NI relative to the GSI of the sham treatment decreased gradually. The GSIs for BA and JA were similar (0.92) and were closer to that for the sham condition. Although the GSI for NI was 0.62 (Fig.4b), it produced significant IV. These results suggest that the same phenotype emerged despite the different profiles of gene expression; this observation reflects the chemical-dependent response and the integrated action of multi-target drugs.

JA+CA shared different GSI with JA, CA and NI (0.57, 0.68, 0.93), respectively. So as to GSI exist in JA+BA with JA, BA and NI (0.81, 0.79, and 0.91) (Fig.4b). This suggests that GSI provides an approach independent of ROG and RNOG to represent the pattern of gene expression in chemogenomic profiling. The more overlapping genes in not represent higher GSI, so as to the lower overlapping ones. For example, we observed a higher percentage of overlapping genes (60%) do a lower GSI (0.72) for JA or NI, but a lower percentage of overlapping genes (44%) do a higher GSI (0.93) for JA+CA or NI (44%).

Although usually applied to relatively small numbers of genes[19], hierarchical clustering in an independent analysis showed that JA+CA with NI, JA+BA with CA, and BA with JA, in three different categories (Fig.4c).

PCA1-3 accounted for 80% of the variation (Fig.4e), which was consistent with the clustering among the six groups (Fig.4d). Both methods validated the results of the GSI determined by varying the combination treatment. In independent experiments, the IV and neurological score was respectively significantly smaller ($P = 0.028$) in all groups except the CM group than in the vehicle group (Fig.5a,b). The CIV (Fig.5c) did not differ significantly between groups ($F_{7,127} = 2.68$, $P > 0.05$), but the PIV was smaller in all compound-treated groups ($F_{7,127} = 20.71$, $P < 0.001$).

New morphological features appear predominantly because of modifications of the spatial patterns of gene expression[20], confirming that similar phenotypes are secondary consequences of similar gene expression and that a transcription defect may be crucial to the development of a clinical syndrome. Although each drug profile represents the drug's own signature at the transcriptional and molecular pharmacological levels[21], most of the genes associated with a particular biological function are up- or down-regulated in a similar way, and the conserved functions of groups of genes should be reflected in similar patterns of gene expression in yeast, worms, fruit flies and humans[22]. GSI showed stable variations in multiple comparative studies and indicated a robust association between the GSI shift and plasticity of outcomes. Our results might provide new insights into the

mechanisms of a compound's action in the ischaemic hippocampus that underlie pharmacological plasticity. We believe that systematic drug-design strategies should be directly against multiple targets, and that this novel drug-design paradigm might help develop more efficient compounds than the currently favoured single-target drugs[12], which interactions of the most promising candidates[23] appear to be fundamental to improving future stroke treatment. Integrating the clinical data from a patient's records[24] and other clinical or experimental variables[25] is also promising. Thus, systems to augment expression analysis with automated literature extraction or organization[26] are likely to prove valuable in drawing meaningful and reproducible conclusions. Our data demonstrated a molecular GSI of "pattern signature" in the mouse ischaemic hippocampus that was robustly associated with the pharmacological effects and provided quantitative assessment of a wide range of responses. These results support the idea that functional predictions based on molecular phenotype association[27] provide a guide for studying combinations in system-oriented drug design[9]. Developing innovative scientific methods for discovery, validation, characterization and standardization of multi-component botanical therapeutics emerging synergic outcomes, and the network and pathways is essential[28].

References and Notes:

1. H.Li, M.Zhan.*Bioinformatics* 22, 96(2006)

2. I. Hovatta, et al. *Nature* 438,662(2005)

3. C.W.Whitfield, A.M.Cziko, G.E.Robinson.*Science* 302, 296(2003)

4. M. Iida1, et al. *Carcinogenesis* 24, 757(2003)

5. J.Lamb, et al. *Science* 313, 1929(2006)

6. C.T.Workman, et al. *Science* 312, 1054(2006)

7. I. Poola, et al. *Nature Medicine* 11, 481(2005)

8. R.Balasubramaniyan, et al. *Bioinformatics* 21, 1069(2005)

9. H.Kitano. Nature Reviews. *Drug Discovery*.6，202（2007）

10. L.Cabusora, et al.*Bioinformatics* 21, 2898(2005)

11. T. Casci.Gene networks go global. *Nature Reviews Genetics* 5, 84 (2004)

12. P.Csermely, et al. *TRENDS Pharmacol.Sci.* 26,178(2005)

13. L.B. Goldstein. *Arch Neural.* 55, 454(1999)

14. M.F. Cano-Abad, et al. *J. Biol. Chem*. 276, 39695(2001)

15. D. Lie, et al. *Nature* 437, 1370(2005)

16. F. Troglio, et al. *Proc. Natl Acad. Sci. USA* 101, 15476(2004)

17. D.J. Gladstone, S. E. Black, A. M. Hakim. *Stroke* 33, 2123(2002)

18. I. Hovatta, et al. *Nature* 438,662(2005)

19. C. H. Chung, P. S. Bernard, C. M. Perou. *Nature Genet.* 32 (suppl.), 533 (2002).

20. N.Gompel, et al. *Nature* 433, 481(2005)

21. M.Bredel, E.Jacoby.*Nat.Rev.Genet.*5, 262(2004)

22. J.Quackenbush.*Science* 302, 240(2003)

23. S. J. Hong, et al. *Proc. Natl Acad. Sci. USA* 101, 2145(2004)

24. J.D.Potter. *Nature Rev. Genet.* 2, 142(2001)

25. A.J.Butte, et al. *Proc. Natl Acad. Sci. USA* 97, 12182(2000)

26. T.Ideker, et al. *Bioinformatics* 18 Suppl 1, S233 (2002).

27. R.B.Stoughton, S.H.Friend. *Nat.Rev.Drug Discovery*.4, 345(2005)

28. B.M.Schmidt,et al. *Nature Chemi.Bio* 3, 360(2007)

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature

**Acknowledgements** We thank Prof.Yongming Wang and Xuejun Zhang of Tianjin Medical University for help to MRI analysis, MD. Yuankai Fu of Human Genomic Research Center of China for cDNA microarray spotting and analysis in Genespring software. We thank Prof. Qiguang Cheng of Southeast University of China for help to biological statistic design and Prof. Weibo Lu of China Academy of TCM for prepares manuscript of this paper also.

**Author Information** The authors declare competing financial interests, details accompany the full-text HTML version of the paper at www.nature.com/nature.Correspondence and requests for materials should be addressed to Zhong Wang (zhonw@vip.sina.com).

Figure 1. Overview of the systemic analysis line. **a**-**c**, Top-down approach. **d**-**e**, Bottom up method. StkE, Science Signal Transduction Knowledge Environment;MAPK, mitogen-activated protein kinase; PCA, principal component analysis;GSI, global similarity index; IV,Infarction volume; PIV, peripheral IV; CIV, central IV; IE,Indenpent experiment; RT-PCR, Reverse Transcription-Polymerase Chain Reaction.



Figure 2 Plasticity of pharmacological phenotype and compound-dependent altered genes in the ischemic mouse hippocampus. **a**, Plasticity of IV in mouse hippocampus with sham, vehicle, BA, JA, CA, CM or NI treatment. $*P < 0.05$, $**P < 0.01$ vs. vehicle. **b**,Altered genes in the ischemic mouse hippocampus in animals treated with vehicle, CM, BA, JA, CA, or NI are indicated as a function of both the fold difference and statistical significance (*P*) for each of the 374 cDNAs on the microarray (tabulated *P*-values from ANOVA, $n \geq 9$ animals per group). Statistical analysis was performed on the mean expression levels of all 374 cDNAs in each of six experimental groups. Numbers represent different (in italic) or overlapping genes in two or more groups. **c**, Different contributing genes in the direct comparison analysis and PCA. Eight overlapping genes existed in all pharmacologically significant groups compared with the vehicle group. The direct comparison model

showed that expression of three genes decreased (in panel) after combination therapy (above). The magnitude of expression differences is shown as the base average ratio of the gene expression level in the treated mouse ischemic hippocampus relative to that in the vehicle-treated ischemic hippocampus. Higher expression in the hippocampus is shown in red and lower expression in blue; the colour intensity is proportional to the magnitude of the expression difference as indicated by the colour bar at the bottom of the figure. **d**, Overlapping genes in all compound-treated scaling groups. **e**, Selected gene expression (Rgs6,regulator of G-protein signaling 6) was validated in real-time PCR. Kcnmb, large-conductance calcium-activated; Tcf, T cell factor;MMP, matrix metalloproteinase;Dgke, iacylglycerol kinase, epsilon; Camk, calcium/calmodulin-dependent proten kinase; Eef. elongation factor; Selenbp2, Mus musculus selenium binding protein 2; Calm, calmodulin; Cbx, chromobox homolog;Grb, growth factor receptor bound protein

Figure 3 Molecular profiling shifts in combination therapy comparing to that of single compound. **a**, The numbers of altered genes in ischemic mouse hippocampus treated with compounds alone (JA, CA) or in combination (JA+CA) and the combination of JA+BA relative to treatment with JA or BA.

**c**, Numbers indicate different and overlapping genes in hippocampus treated with a single compound or with combined compounds. Visualizing gene expression of JA, CA and two-compound treatment, compare NI (**b**) and JA+BA with single compounds (**d**). Only the 374 cDNAs exhibiting > 1.25-fold mean difference between compounds and vehicle are shown. **e**, The gene expression profiles of all groups contrast in nine pathways. The 374 cDNAs predicted the outcome in all effective groups and contribute to many pathways, such as Wnt, p53, MAPK and GPCR (arbitrary fold criterion depicted for graphic representation only). Red denotes an increased mRNA level compared with the average of all animals. Green denotes a decreased mRNA level. **f**, Erk-MAPK networks display in different groups. Only the gene altered significantly in all groups and which of the Pearson coefficients > 0.5 were selected and linked with a line. White, plum and cardinal red circles represent altered genes whose ratios exceeded 1.5, 1.7 and 2.0, respectively. Green circles represent genes changed significantly by the combination treatment but not by the single treatment; blue circles are the genes only changed significantly by the single treatment.

Figure 4 GSI of compound-oriented profiles of gene expression and validation by an independent experiment. **a**, Our global similarity analysis

approach compared with the Euclidean distance and Pearson coefficient. **b**, GSI and percent of overlapping and non-overlapping genes in the two groups. **c**, Hierarchical clustering and PCA indicate the existence of three categories of compound-dependent gene expression profiles(**d**). **e**, PCA indicated similar results of clustering analysis for the PC1, PC2 and PC3 maps; the sum accounted for 81.08% of the variance in the data and the individual contributions were 57.63%, 13.15% and 10.30%, respectively.

Figure 5 Gene expression profiles predict pharmacological outcomes. IV was significantly lower after treatment with all compounds except CM. **a**, IV. F =16.23, ** $P < 0.01$ vs. vehicle; mean ± SD, ANOVA, n = 10).**b**, Behaviour score (F = 12.34, ** $P < 0.01$ vs. vehicle; mean ± SD, ANOVA, n = 10). JA+CA and NI produced the highest scores. **c**, Magnetic resonance imaging (MRI) results were consistent with the behaviour scores in the PIV but not CIV. This pharmacological effect confirmed the categories based on gene expression profiles of the ischemic mouse hippocampus *, #$P < 0.05$, **, ##$P < 0.01$ vs. vehicle.

## Methods Summary

### GSI Approach

The similarity between chips can be calculated by pairing comparing gene expression profile. For two gene microarray $a$ and $b$, containing $n$ genes, the gene expression vector is $x=[x_1,x_2,\ldots,x_i,\ldots,x_n]^t$ and $y=[y_1,y_2,\ldots,y_j,\ldots,y_n]^t$ respectively. The similarity between them can be calculated by cosine coefficient

$$similarity=\frac{x\cdot y}{|x|\cdot|y|} \quad (1)$$

Gene microarray can simultaneously detect expression of thousands of genes. For two gene expression vector $x$ and $y$, the significant expression gene is $Gene_x$ and $Gene_y$ respectively. In addition, the total expression gene is the union of $Gene_x$ and $Gene_y$. The number of total expression gene is $n_{Change\_gene}$. The number of discrepancy genes is $n_{diff}$, which means there are $n_{diff}$ genes significantly expressed in microarray $a$ instead of $b$ or in the inverse situation. We calculated the similarity between two microarray data as

$$Similarity_{chip}=Sim_{total\_gene}-(1-Sim_{diff\_gene})\times n_{diff}/n_{Change\_gene} \quad (2)$$

Here, $Sim_{total\_gene}$ and $Sim_{diff\_gene}$ is the similarity of original gene data and the discrepancy genes according to formula (1). $n_{diff}$ In addition, $n_{Change\_gene}$

is the number of discrepancy gene and all changed genes, respectively. The second half part of formula (2) is modification to the similarity of all changed genes.

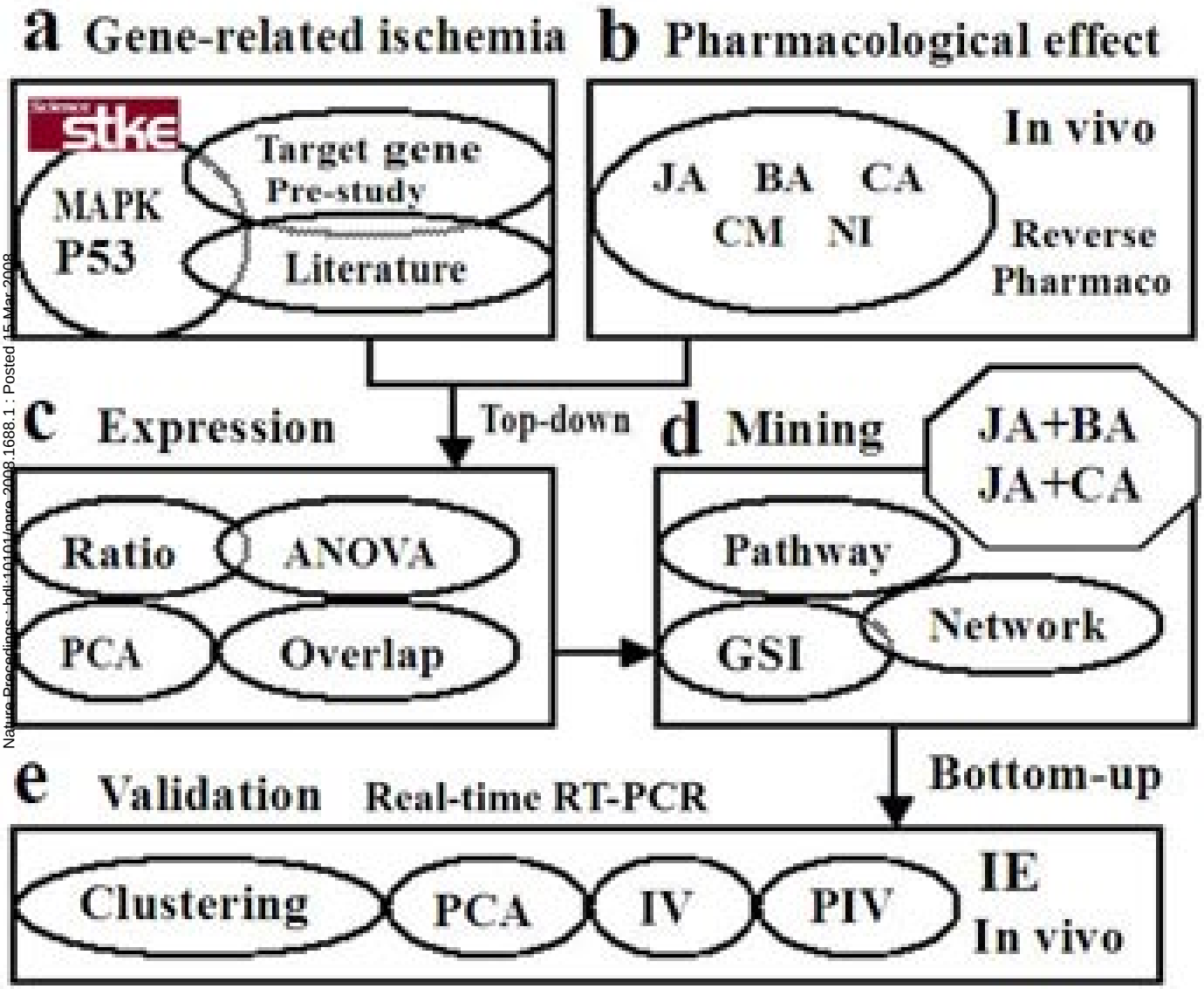
The similarity can be used to evaluate the consistency of data from different blocks in one chip. Because certain ratio value of one gene maybe losts in the practice. We proposed a wrong data coefficient to describe those situations. It is calculated by

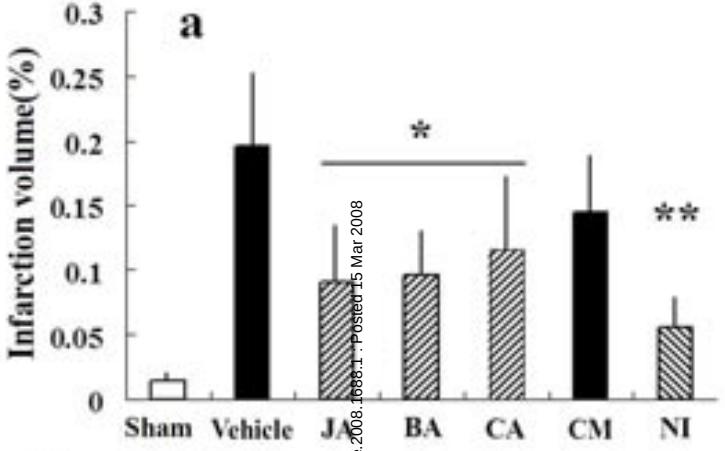
$$wrong\_coef = 1 - m/n \quad (3)$$

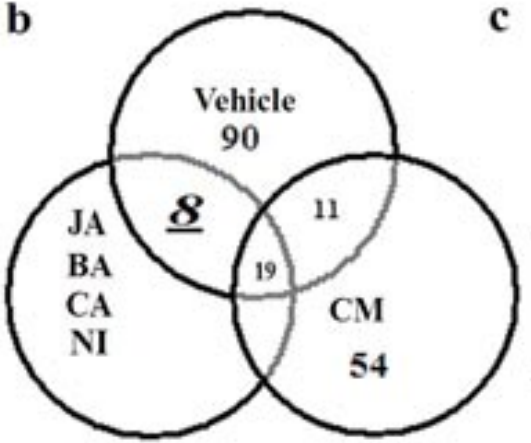
Here, $m$ is the number of absent data and $n$ is the number of all genes in one microarray. Moreover, the similarity between different blocks is calculated by

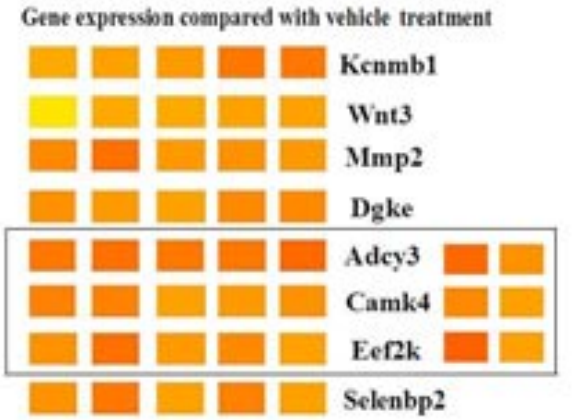
$$Similarity_{block} = Wrong\_coef \times (Sim_{total\_gene} - (1 - Sim_{diff\_gene}) \times n_{diff} / n_{Change\_gene}) \quad (4)$$

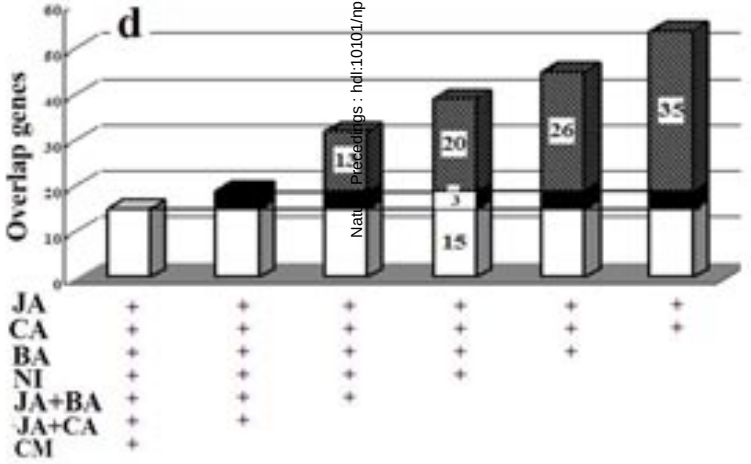
The similarity between samples from same group can be calculated in the same manner. In addition, it can be used to describe the magnitude of biological variation. Moreover, the similarity of microarray data from different groups can also be calculated according to formula (2) to observe whether they have similar gene expression profile. Here, we used GSI to represent $Similarity_{chip}$ only and data of $Similarity_{block}$ were not shown.

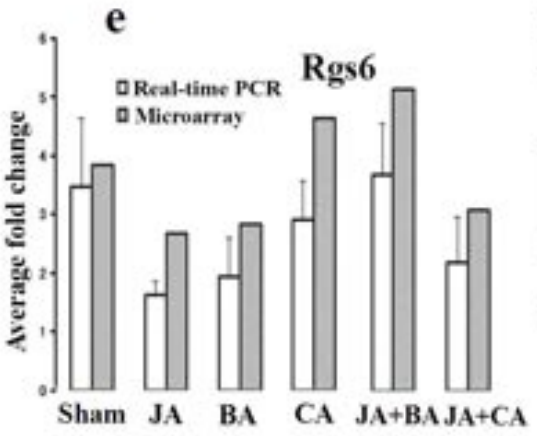


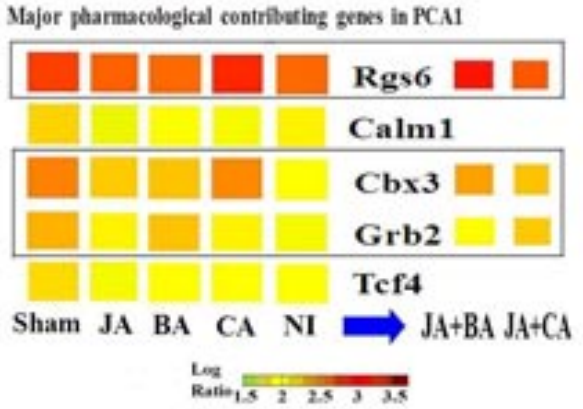


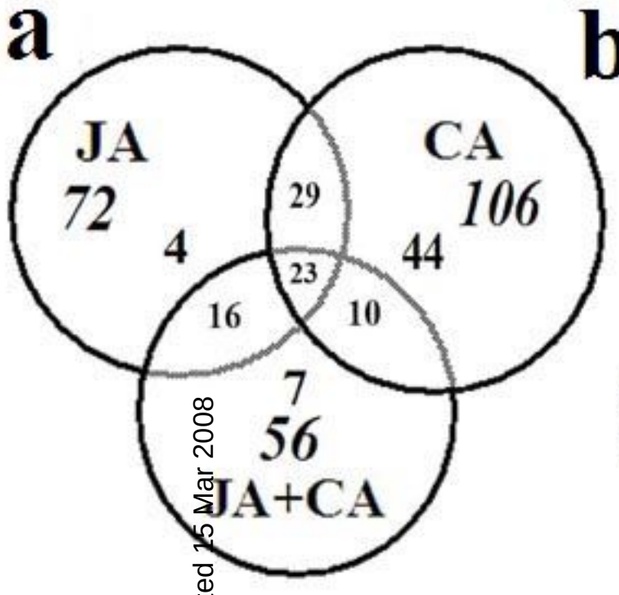



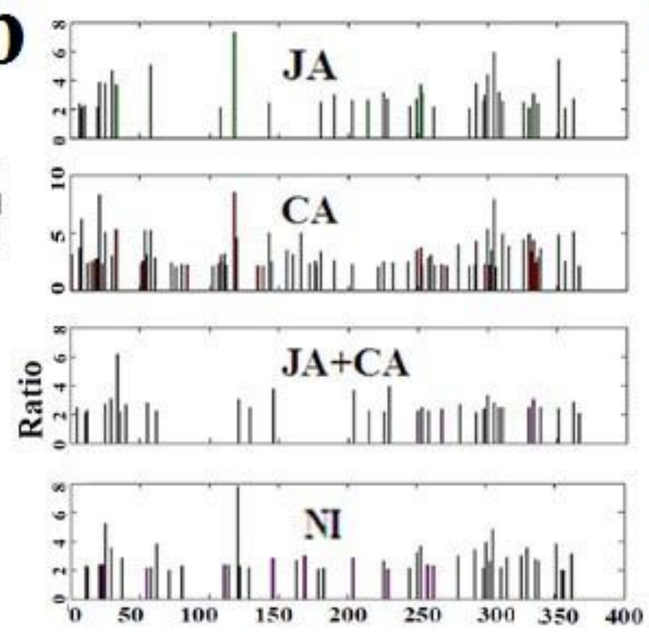



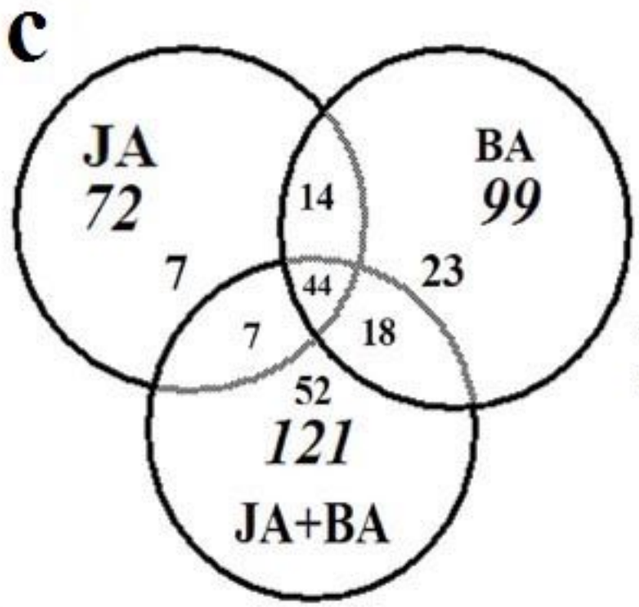



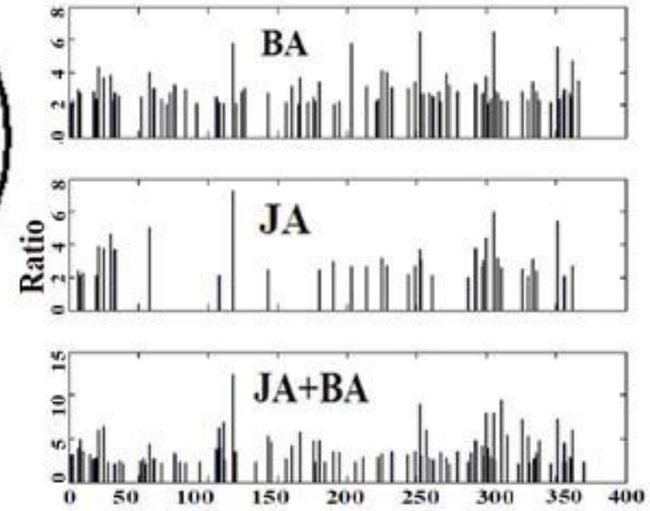



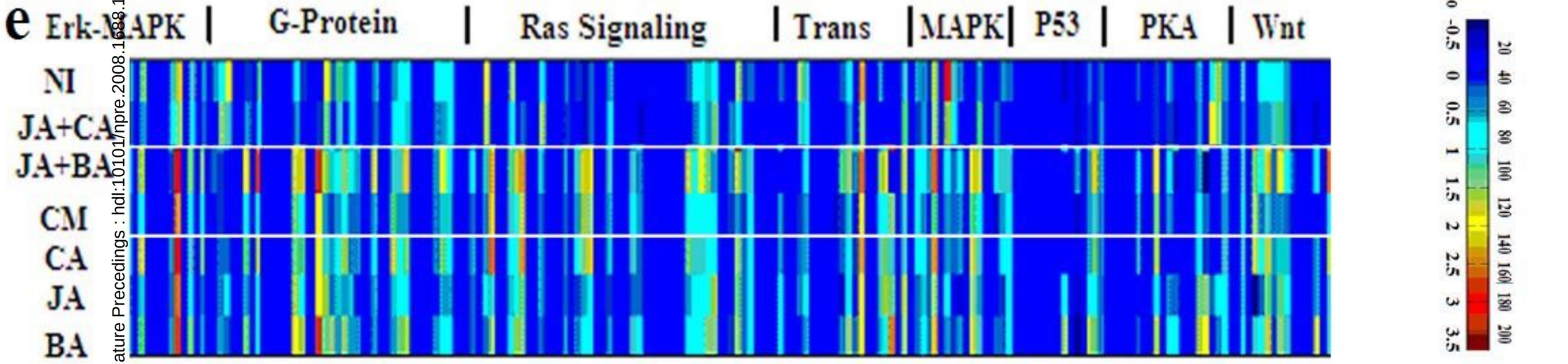



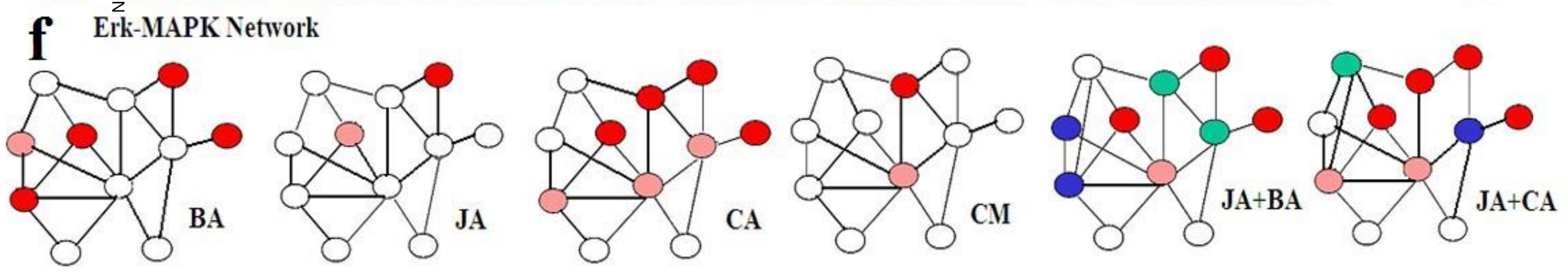


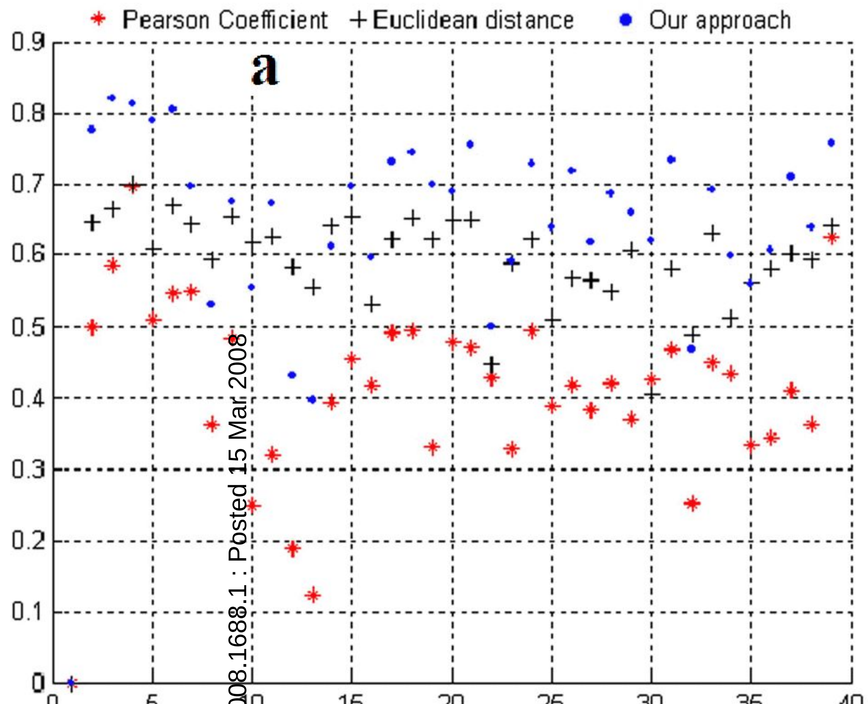



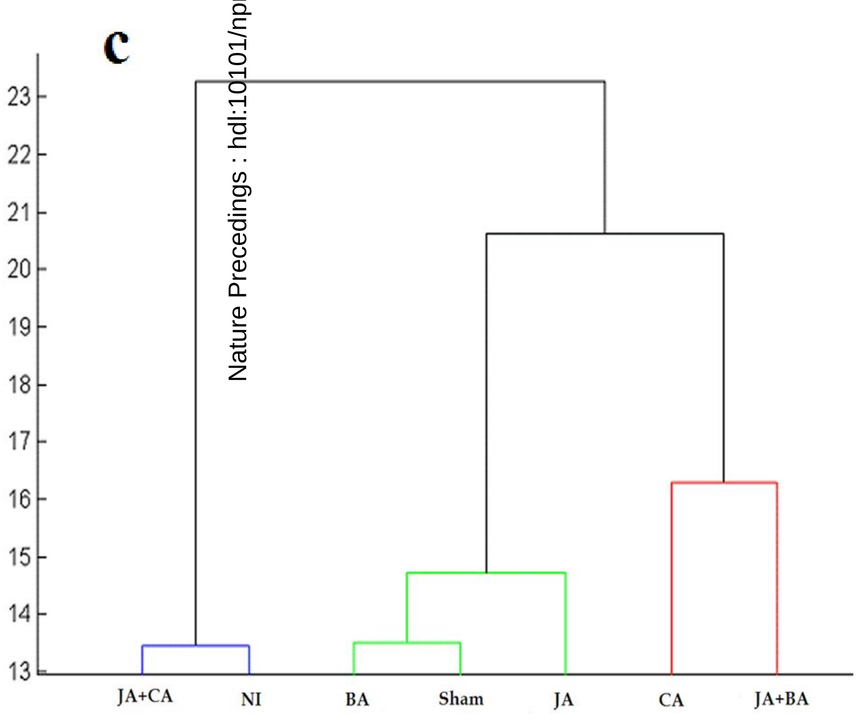


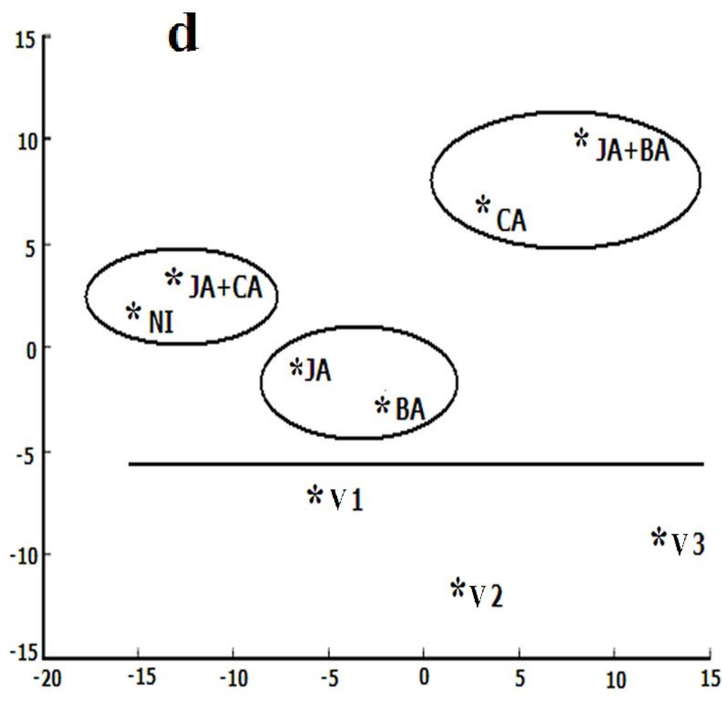

**b**

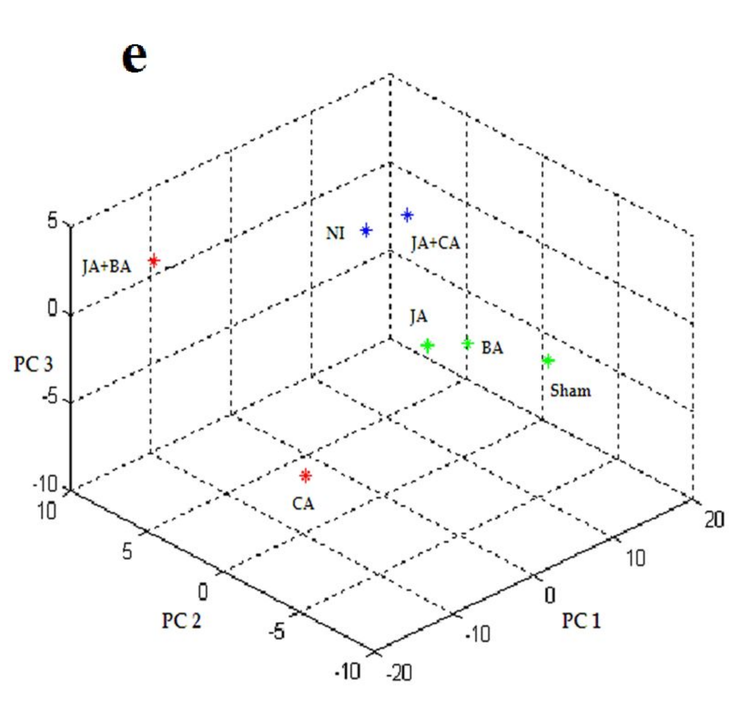
| | Sham | JA | BA | CA | NI | JA+CA | JA+BA |
|---|---|---|---|---|---|---|---|
| Sham | | 0.87 | 0.92 | 0.74 | 0.62 | 0.61 | 0.64 |
| JA | 53/56(49%) | | 0.92 | 0.86 | 0.72 | 0.68 | 0.81 |
| BA | 56/77(42%) | 58/55(51%) | | 0.82 | 0.66 | 0.64 | 0.79 |
| CA | 52/92(36%) | 52/74(41%) | 67/71(49%) | | 0.64 | 0.57 | 0.91 |
| NI | 43/66(39%) | 50/34(60%) | 51/59(46%) | 50/68(42%) | | 0.93 | 0.59 |
| JA+CA | 38/70(35%) | 39/50(44%) | 43/69(38%) | 33/96(26%) | 36/46(44%) | | 0.54 |
| JA+BA | 50/111(31%) | 51/91(36%) | 62/96(39%) | 84/63(57%) | 47/89(35%) | 35/107(25%) | |

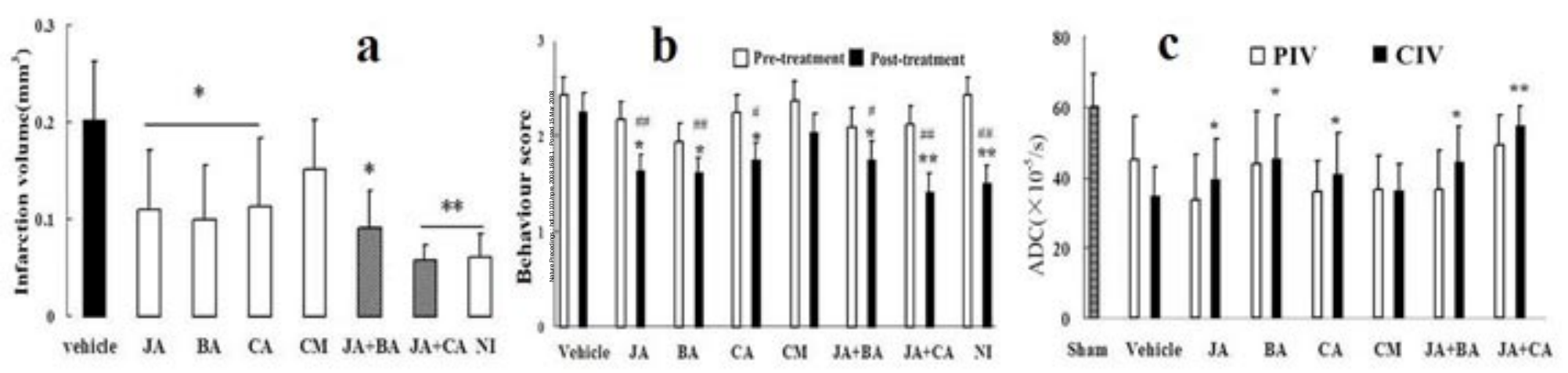
Overlap/non-overlap genes (Overlap genes percent)






a
Infarction volume(mm³)
0.3
0.2
0.1
0
*
*
**
vehicle
JA
BA
CA
CM
JA+BA
JA+CA
NI
b
Pre-treatment
Post-treatment
Behaviour score
3
2
1
0
Vehicle
JA
BA
CA
CM
JA+BA
JA+CA
NI
c
PIV
CIV
ADC(×10⁻⁵/s)
80
60
40
20
0
Sham
Vehicle
JA
BA
CA
CM
JA+BA
JA+CA